Feature Selection Method EigenNet in High Dimensional Space

Rahul Barikeri (rahulravindr), MSCS
Ashutosh Modi (modia), MSCS
Rubein Shaikh (rubein), MSCS
Department of Computer Science and Engineering
University of South Florida
Tampa, Florida 33620

***Abstract*— It has always been focus to retain as much information as possible by removing redundant features while learning a predictive model. And when data is huge with large number of features computation gets expensive in high dimensionality. To ameliorate the curse of dimensionality feature selection is vital technique in Machine Learning which gives subset of features that to with improved prediction performance, reduction in computation time and gives better understanding of data. An irrelevant feature when combined with other may give useful information. On the other hand important features combined may also loose information. This [7] paper explains why the best two independent features may not be the two best. Moreover, it is very challenging to classify data when number of examples is less than number of features. Selecting features in these cases reduces the probability of overfitting. So it is very important which algorithm we choose. In this project we studied different feature selection techniques and how EigenNet method works for high dimensional data.**

1. Details

Feature selection and feature extraction are different techniques and used for different purposes. Feature selection is the process of creating subset out of original features. For checking performance simplest way is to create all possible subset calculate the error. Set with minimum error will be the best choice. But to do so it is very cumbersome on data with large number of features. This type of evaluation and others are generalized into three categories mainly:

- Wrapper method: use predictive model to calculate error rate (score)
- Filter Method: use statistical measures to rank features instead of calculating error rate
- Embedded Method: are kind of built into learning algorithm

Some of the feature selection techniques are:

- Correlation
- Information Gain
- Maximum-relevance-minimal-redundancy (mRMR)
- Relief Score
- EigenNet

2. Literature Survey

2.1. Why the topic of feature selection?

For the past 2 decades dimensionality of the data involved in Data Mining has increased exponentially. And for existing learning algorithms it is very challenging to implement for large number of features.

We have been following different feature selection methods and their applications mainly in medical domain. One of the novel methods we found is EigenNet which is a hybrid of generative and conditional model. Where in conditional model learns classifier and selects eigen vectors and generative captures correlation leading better estimation. So idea is to guide feature selection using eigenstructure.

2.2. Lasso:

It is a Least Absolute Selection and Shrinkage Operator and uses L1 norm.

- Pros:
  - For high dimension space coefficients tend to 0
  - Sparsification of $\overrightarrow{\beta}$
  - Which feature has real impact select them
- Cons:
  - Fails to perform group selection
  - Total # of feature selection is bounded by # of instances

2.3. Ridge Regression:

It uses L2 norm.

- Pros: For correlated features puts constraint on their weights
- Cons: Can not perform feature selection unless $\lambda \to \infty$

2.4. Elastic Net:

It uses both L1 and L2 norm.

- Pros:
  - For high dimensional data correlation between features can be high
  - Correlated variable sometimes form a group
  - If one is selected from this group we would want to select entire group
- Cons: Does not exploit correlation information

2.5. EigenNet:

Is a hybrid of conditional and generative model.

- Conditional model
  - Learns classifier → selecting eigen vectors
- Generative model
  - captures correlation → estimation
- Eigenstructure guiding variable selection

2.6. Data Sets Used:

Microarray dataset [6]: http://csse.szu.edu.cn/staff/zhuzx/Datasets.html

1) ALL-AML
   Acute Lymphoblastic Leukemia (ALL)
   Acute Myeloblastic Leukemia (AML)
2) Ovarian

Both are used for classification task. and eatures represent mainly genes.

Type	**Property**	
Database	Leukemia (ALL-AML)	Ovarian
Number of Attributes	7129	15154
Number of Instances	72	253
Classes	47 (ALL)	91 (Normal)
	25 (AML)	162 (Cancer)

TABLE 1: Data Set Properties

2.7. Tools Used:

- Weka
- RStudio
- Matlab

Algorithm 1 Elastic Net

```
//START
1. Load required libraries and dataset
2. Divide training data in two sets
3. For different values of alpha
   a. Use cross validation and learn model
   b. Plot graph for CV
   c. Make predictions
   c. Calculate accuracy
   d. Plot ROC curve
//END
```

3. Implementation and observations

We checked accuracy for our selected dataset with different algorithms mainly elastic net, decision tree and SVM as we could not implement the EigenNet in given time constraint. But we used RStudio for elastic net and Weka to compare with other classifiers. We used glmnet [8] library for tuning α and λ parameters. We used logistic regression and following is our objective function.

$$\min_{\beta\,\lambda} J(\beta) = \frac{1}{2N}\sum_{i=1}^{N}\left(f_\beta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\|\beta\|_p$$

We also tried writing subroutines in MATLAB for elastic net but could not tune our model.

3.1. RStudio

Algorithm 1 is the pseudo code of our implementation in R script.

For dividing data in two sets we used random sampling without replacement. 70% of data used for training and 30% for testing. glmnet package was used to learn model. We ran for loop for different values of α from 0 to 1 with steps of 0.1.

Following are the result we got on RStudio.

3.2. Results

Table 2 gives results we got. Where values represents AUC and accuracy percentage are within parenthesis.

	Ovarian	Leukemia
Decision Tree	0.957 (95%)	0.784 (83%)
LibSVM	0.819 (87%)	0.5 (65%)
ElasticNet	1 (100%)	0.97 (96%)

TABLE 2: Results

Figure 1 on the following page represents ROC curve for leukemia and figure 2a on the next page and 2b on the following page are CV plots for Leukemia and Ovarian respectively.

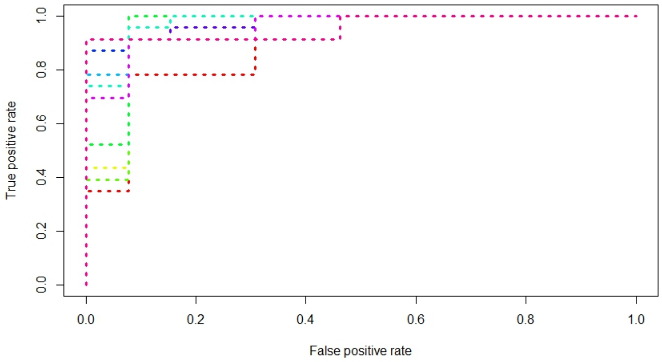
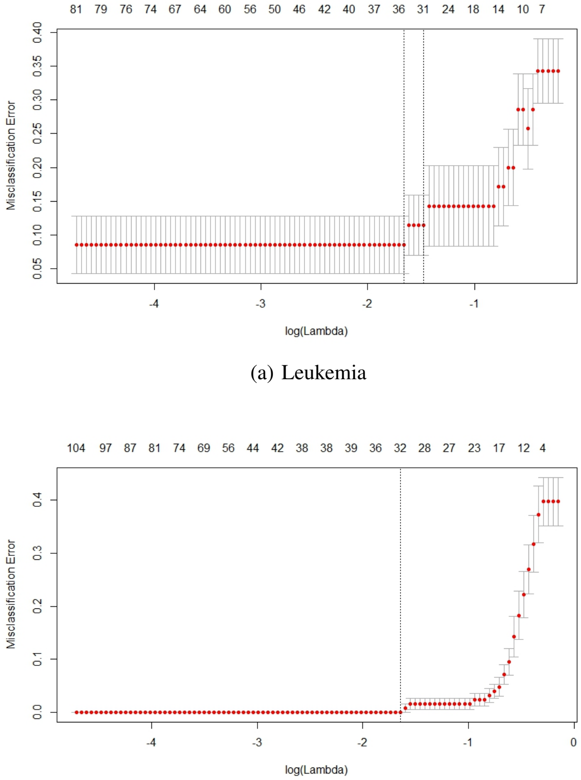


Figure 1: ROC curve - Leukemia



(a) Leukemia

(b) Ovarian

Figure 2: CV Plot

4. Team Efforts:

1) Rahul: Literature Survey and Weka implementation. Review of report and presentation.
2) Ashutosh: Literature Survey, MATLAB and RStudio implementation. Report and presentation creation.
3) Rubein: Study of different tools i.e. Weka implementation, SAS miner etc. Review of report and presentation.

5. Conclusion

Elastic net performed very well on the selected dataset. High test data accuracy goes in accordance with the low cross validation error. But this may be because of the small data set. Moreover, we have used 10 fold cross validation and have not compared with other methods. For large data set, comparable to number of features, performing 10 fold CV would be time expensive. Also, implementing EigenNet remains open as a future work.

References

[1] Qi, Yuan, and Feng Yan. *EigenNet: A Bayesian hybrid of generative and conditional models for sparse learning.* arXiv preprint arXiv:1102.0836 (2011).

[2] Shardlow, Matthew. *An analysis of feature selection techniques.* The University of Manchester (2016).

[3] Refaeilzadeh, Payam, Lei Tang, and Huan Liu. *On comparison of feature selection algorithms.* Proceedings of AAAI workshop on evaluation methods for machine learning II. 2007.

[4] Chandrashekar, Girish, and Ferat Sahin. *A survey on feature selection methods.* Computers & Electrical Engineering 40.1 (2014): 16-28.

[5] Guyon, Isabelle, and André Elisseeff. *An introduction to variable and feature selection.* Journal of machine learning research 3.Mar (2003): 1157-1182.

[6] Zexuan Zhu, Y. S. Ong and M. Dash, “Markov Blanket-Embedded Genetic Algorithm for Gene Selection”, Pattern Recognition, Vol. 49, No. 11, 3236-3248, 2007.

[7] Thomas M. Cover. *The Best Two Independent Measurements are Not the Two Best.* IEEE Transactions on Systems, Man and Cybernetics, SMC-4(1):116–117, January 1974.

[8] https://cran.r-project.org/web/packages/glmnet/index.html